View-Spectral Reconciliation Learning for Text-based Multispectral Aerial–Ground Person Re-Identification

Zhongao Zhou[1,2*] **Bin Yang**[2*] **Yuxuan Zhao**[2] **Yudi Xie**[2] **Jun Chen**[1,2†] **Mang Ye**[1†]
[1] National Engineering Research Center for Multimedia Software, Wuhan, China
[2] School of Computer Science, Wuhan University

Abstract

Conventional text-based person re-identification (ReID) approaches mainly rely on visible-spectrum images captured from ground-level viewpoints. However, with the increasing prevalence of UAVs and multispectral imaging systems, pedestrians can be captured from both ground and aerial viewpoints, and different spectral sensors provide complementary cues that reveal diverse pedestrian characteristics. As a result, conventional text-based ReID methods become inadequate for such retrieval scenarios. To address this limitation, we introduce a novel task termed Text-based Multispectral Aerial-Ground Person Re-Identification (**TMAG-ReID**), which aims to retrieve pedestrian images across aerial and ground viewpoints under multispectral scenarios using textual descriptions. To facilitate this task, we construct the **TMAG** dataset, which consists of multispectral pedestrian images captured from aerial and ground viewpoints, paired with corresponding textual descriptions. Furthermore, we propose a novel **View-Spectral Reconciliation Learning framework (VSRL)** that leverages textual guidance to interweave discriminative semantics across distinct views and spectral modalities. This fusion yields robust cross-view and cross-modal pedestrian features, thereby enabling high-performance retrieval. Extensive experiments on the TMAG dataset demonstrate the effectiveness of VSRL. The code is available at https://github.com/ZhouZhongao/VSRL.

1 Introduction

Text-based Person Re-identification (T-ReID) [19, 1, 2, 6, 61] aims to retrieve a target pedestrian image from a gallery according to textual descriptions. Compared with traditional image queries [44, 47, 22, 32, 49, 53, 41, 7, 14, 54, 8, 9], textual descriptions offer greater flexibility and are more readily accessible. Consequently, T-ReID has gained growing research interest, extending the applicability of ReID systems [11, 58, 51, 26, 50, 5, 46, 28, 38, 17, 62, 63, 13, 23, 33, 56, 10, 45] toward more practical scenarios.

With the rapid development of intelligent technologies, unmanned aerial vehicles (UAVs) [21, 20, 34, 48, 56] and multispectral imaging devices [27] have been widely adopted in visual tasks. UAVs offer high mobility and broad coverage, enabling flexible data acquisition over large areas. Separately, multispectral sensors provide complementary information across visible (RGB), near-infrared (NIR), and thermal-infrared (TIR) bands, allowing for robust capture of high-quality visual data under challenging environments. However, existing T-ReID datasets [19, 6, 61] are largely restricted to single-view ground RGB images. This limitation raises ***Challenge 1****: Can we propose a new task and construct a dedicated dataset for text-based multispectral ground-aerial person re-identification?*

Pedestrian images captured by UAVs often suffer from smaller scales, pronounced geometric distortions, and substantial viewpoint variations, leading to a considerable appearance discrepancy between

*Equal contribution.
†Corresponding Author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

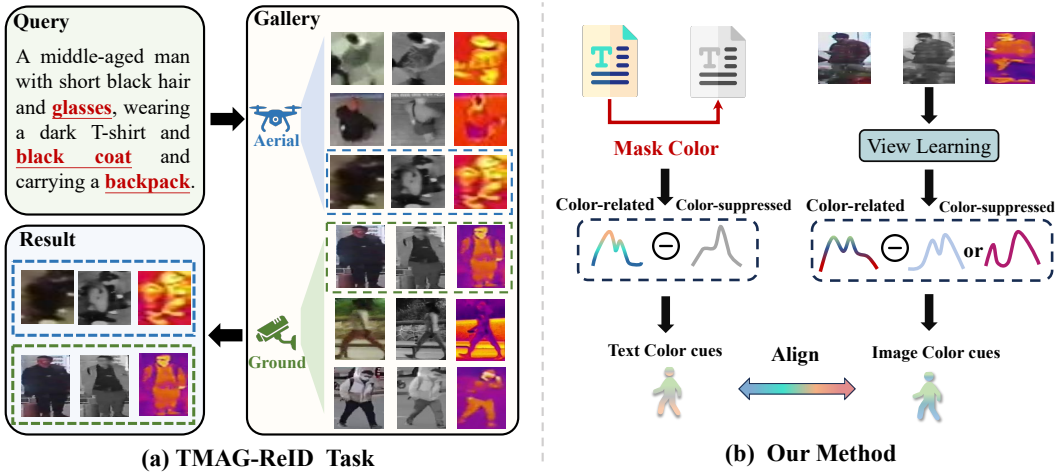

Figure 1: Illustration of the TMAG-ReID task and our method. (a) Given a textual description as a query, TMAG-ReID aims to retrieve the target pedestrian in tri-spectral pedestrian images from the gallery. (b) Cross-modal alignment is guided by anchors derived from color-related semantics in textual features and cross-spectral color discrepancy information.

aerial and ground views. At the same time, textual descriptions are usually grounded in human-visible RGB observations, often from ground-level viewpoints, which makes direct retrieval from aerial imagery challenging. This raises ***Challenge II***: *How to learn viewpoint-invariant representations that effectively bridge the discrepancy between pedestrian appearances and textual descriptions?*

Moreover, textual descriptions predominantly depict attributes observable in the RGB modality, especially human-perceived color cues, whereas NIR and TIR images provide color-weak structural and thermal cues with markedly different visual distributions. Directly matching these RGB-grounded descriptions with NIR/TIR data fails to fully exploit the complementary information contained in these spectra, thereby degrading retrieval performance. This introduces ***Challenge III***: *How can we effectively exploit complementary spectral information while maintaining semantic consistency between RGB-grounded text and color-weak NIR/TIR images?*

To address these challenges, we propose the following contributions. For ***Challenge I***, we introduce a new task, Text-based Multispectral Aerial-Ground Person Re-Identification (TMAG-ReID). As illustrated in Figure 1, given a textual description of a pedestrian, the goal is to retrieve the corresponding tri-spectral pedestrian images from a gallery containing both aerial and ground views. To support this task, we construct a novel dataset TMAG, in which pedestrian images are captured by both ground cameras and aerial UAVs equipped with tri-spectral (RGB, NIR, TIR) sensors. Corresponding textual descriptions are first generated by a multimodal large language model (MLLM) and subsequently refined through manual verification to ensure reliability. For ***Challenge II***, we propose a Bidirectional View Learning Strategy (BVLS), which decouples intra-modal view features from global identity representations while enhancing cross-modal view feature correlations, thus encouraging the model to learn view-invariant and robust pedestrian representations. For ***Challenge III***, we propose Text-driven Multispectral Synergy Learning (TMSL), which formulates text-guided multispectral learning as a text-spectral residual closure problem. By deriving a counterfactual textual color residual and injecting it into color-weak spectra, TMSL encourages the compensated NIR/TIR branches to agree with RGB without directly matching noisy spectral residuals. The contributions are summarized as:

❶ To address multiview-multispectral scenarios, we propose a novel task TMAG-ReID which aims to retrieve tri-spectral images of target pedestrians from arbitrary viewpoints based on textual descriptions. We further construct a new dataset TMAG to support the research on TMAG-ReID.

❷ To address the challenges in TMAG-ReID, we design a new framework VSRL, which consists of Bidirectional View Learning Strategy (BVLS) and Text-driven Multispectral Synergy Learning (TMSL). BVLS learns view-invariant pedestrian representations, while TMSL reconciles text and spectra by compensating color-weak branches with counterfactual textual color residuals.

❸ Extensive experiments on the TMAG dataset demonstrate the effectiveness of VSRL, achieving consistent improvements compared to existing T-ReID methods.

2 Related Work

2.1 Text-based Person Re-identification

Text-based person re-identification (ReID) is first introduced by Li *et al.*[19], who establish the CUHK-PEDES benchmark dataset. Jiang *et al.*[15] implicitly exploit fine-grained interactions to refine the global text-image alignment, achieving performance gains without relying on additional supervision. Chen *et al.*[4] propose a unified framework that enables retrieval of RGB images using arbitrary combinations of text and sketch modalities. Cao *et al.*[3] conduct a comprehensive empirical study, contributing a straightforward and strong TBPS-CLIP baseline for text-based ReID. Qin *et al.*[29] introduce the RDE method to address misaligned samples in ReID, employing Gaussian Mixture Models (GMM) to predict sample alignment and select optimal training samples. However, the aforementioned text-based ReID methods focus primarily on ground-view images, which limits their applicability in practical aerial-ground retrieval scenarios. Therefore, we propose the TMAG-ReID task, which better reflects real-world scenarios. Furthermore, to support the investigation of the TMAG-ReID task, we construct the TMAG dataset which incorporates multiple viewpoints.

2.2 Multispectral Person Re-identification

Zheng *et al.* [57] propose the first tri-spectral ReID dataset RGBNT201 and employ a hierarchical fusion strategy to combine RGB, NIR, and TIR features. Wang *et al.* [40] propose a novel token permutation strategy to strengthen inter-modal interactions and ensure effective multispectral feature alignment. In the EDITOR framework, Zhang *et al.* [49] introduce spatial-frequency masks to refine multi-modal features. Wang *et al.* [42] utilize a mixture-of-experts (MOE) mechanism to adaptively balance decoupled features, thereby improving the robustness of multi-modal representations. Existing multispectral ReID research primarily focuses on image-to-image retrieval, which fails to address retrieval scenarios where the image query is missing. Additionally, all images are captured from a single ground-level viewpoint, which does not align with the multi-view retrieval scenarios encountered in real-world applications. The TMAG-ReID task and the TMAG dataset proposed in this paper significantly extend the practicality of ReID.

3 Dataset

Existing text-based or multi-view ReID datasets [52, 25, 60] mainly focus on RGB images, while existing multispectral datasets [57] are typically limited to ground-view image-based retrieval without textual descriptions. To address the limitation in the TMAG-ReID task, we construct a new text-based multispectral dataset TMAG which incorporates both aerial and ground views. The TMAG dataset contains 258 identities and 6,656 tri-spectral image groups, comprising 5,070 ground-view and 1,586 aerial-view groups. It is divided into 185 training identities with 5,381 groups and 73 test identities with 1,275 groups. Several samples are shown in Figure 2.

3.1 Multispectral Image Collection

We collect data from five ground cameras and two UAVs, covering both ground and aerial perspectives. The UAVs are DJI Matrice 300 RTK platforms with Zenmuse H20T+ tri-spectral cameras, while the ground systems use customized tri-spectral spherical cameras. UAV images are captured at 3840×2160 or 1920×1080 for RGB/NIR and 640×512 for TIR at 30 fps; ground images are captured at 1920×1080 for RGB/NIR and 640×480 for TIR at 25 fps. All spectral channels are synchronously captured, and UAVs fly at 20–50 meters to provide distinct aerial viewpoints. We retain pedestrians appearing in at least two cameras and filter low-quality samples using NIMA [35].

3.2 Textual Description Generation

Following IDEA [43], we generate textual descriptions with MLLMs using prompts covering attributes such as gender, hairstyle, and age. To reduce ambiguities caused by the small pedestrian

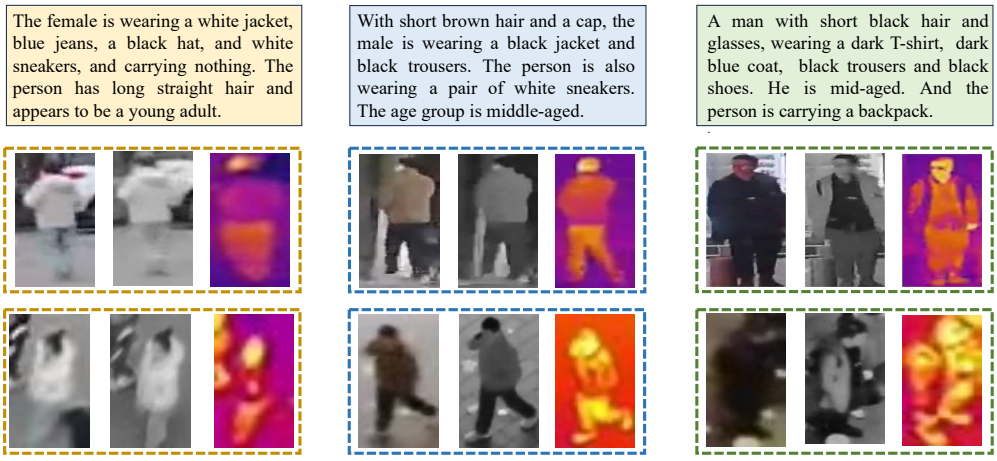

Figure 2: Examples from our TMAG dataset. Each group consists of one textual description and a set of tri-spectral images captured from ground (the second row) and aerial (the third row) viewpoints.

scales in aerial images, we jointly provide the RGB image sequences of each identity to the MLLM. Since pedestrians often appear at small scales in UAV-captured images, we group all sequences of the same ID and feed them jointly into the MLLM for more reliable description generation. Each identity is associated with six verified RGB-grounded descriptions. During evaluation, all six descriptions are used as independent textual queries.

4 Method

4.1 Baseline

Our method utilizes a dual-encoder transformer architecture based on the CLIP model, which provides effective cross-modal alignment capabilities between text and images. As shown in Figure 3, the framework contains a text encoder and an image encoder shared across all visual modalities.

Image Encoder. Given an instance of tri-spectral images $I = \{I^{rgb}, I^{nir}, I^{tir}\}$, each image is cropped into uniformly shaped patches. The image patches are mapped into fixed-dimensional embedding vectors and combined with positional embeddings. A [CLS] token and a view token are appended to the beginning and end of the sequence, respectively, and jointly fed into the model to learn global and view-specific features. We denote the set of spectral branches as $\mathcal{M} = \{rgb, nir, tir\}$. For each $m \in \mathcal{M}$, the image encoder ϕ_v outputs $F_m = \{C_m, F_m^1, F_m^2, ..., F_m^l, V_m\}$. C_m and V_m represent the global [CLS] representation and view token of I^m.

Text Encoder. Given an input text T, we extract the text representation using the CLIP text encoder ϕ_t, which is a Transformer [37] modified by Radford *et al.* [30]. The input text description is tokenized using the Byte-Pair-Encoding (BPE) [31] with a vocabulary size of 49152. Then the tokenized sequence $T = \{T_{sos}, T_1, T_2..., T_s, T_{eos}\}$ is fed into the transformer to capture the contextual relations among tokens. The T_{sos} and T_{eos} tokens represent the start and end of sequence. Finally, the output at C_{txt} from the top transformer layer is projected into the image-text joint embedding space to obtain the global text representation.

Objective Function. To enable text-based retrieval of multispectral pedestrians, our goal is to learn a shared feature space between text and multispectral images. We first sum the global features of the three spectral modalities to obtain a fused tri-spectral visual representation. Then, we optimize the baseline with standard ID loss [59] and SDM loss [15]. The overall loss function is defined as:

$$\mathcal{L}_b = \mathcal{L}_{id} + \mathcal{L}_{sdm}. \tag{1}$$

4.2 Bidirectional View Learning Strategy

To capture viewpoint information embedded in images, we append a view token to the end of the image patches and feed them into the network. To encourage the view token to encode reliable viewpoint

information, we utilize the viewpoint category as a label y_{view} and compute the cross-entropy loss:

$$\mathcal{L}_{view}^{m} = \mathcal{L}_{CE}(z_m, y_{view}), m \in \mathcal{M} \tag{2}$$

where m represents the spectral modality, z_m denotes the logits of V^m. The view loss across all spectral modalities is computed as:

$$\mathcal{L}_{view} = \mathcal{L}_{view}^{rgb} + \mathcal{L}_{view}^{nir} + \mathcal{L}_{view}^{tir}. \tag{3}$$

Then we consider the optimization of the view token from two complementary directions. The global features of pedestrians should be view-invariant, implying that the correlation between the [CLS] token and the view token should be minimized. To reduce view leakage into identity representations, we explicitly penalize the correlation between the global [CLS] token and the view token. Specifically, the two feature representations are constrained to be mutually orthogonal in the latent space. Therefore, we employ the following objective to guide the optimization of view tokens:

$$\mathcal{L}_{BVLS}^{1} = \sum_{m \in \{rgb,nir,tir\}} |\cos(C_m, V_m)|, \tag{4}$$

where $\cos$ and $|\cdot|$ denote the function of cosine similarity and absolute value operators, respectively. C_m and V_m represent the [CLS] token and view token for spectral modality m.

Since the tri-spectral channels are captured synchronously from the same platform, their view labels are shared. We therefore encourage view tokens from different spectra of the same sample to be consistent. Therefore, we introduce another loss function to enforce the consistency of view tokens across spectra:

$$\mathcal{L}_{BVLS}^{2} = \sum_{m,n \in \{rgb,nir,tir\}}^{m \neq n} (1 - \cos(V_m, V_n)) \tag{5}$$

Finally, a Bidirectional View Learning Strategy loss function is used to optimize the view token:

$$\mathcal{L}_{BVLS} = \mathcal{L}_{view} + \mathcal{L}_{BVLS}^{1} + \mathcal{L}_{BVLS}^{2} \tag{6}$$

4.3 Text-driven Multispectral Synergy Learning

Through the bidirectional view learning strategy, the model is encouraged to suppress view-dependent variations in global pedestrian representations. However, another intrinsic asymmetry remains in TMAG-ReID: textual descriptions are mainly grounded in RGB observations, where human-perceived color attributes are highly discriminative, while NIR and TIR images provide color-weak structural and thermal cues. Directly aggregating tri-spectral features may therefore entangle RGB-dominant color semantics with color-suppressed spectral cues. To address this issue, we propose Text-driven Multispectral Synergy Learning (TMSL), which formulates text-guided multispectral learning as a text-spectral residual closure problem.

Motivated by analogical reasoning in language and vision-language embeddings [24, 16, 12], we treat explicit color words in the text as a controllable semantic intervention. Given a textual description T, we introduce a deterministic counterfactual color intervention $\mathcal{I}_c(\cdot)$, which replaces all explicit color words with the textual placeholder $[\mathrm{MASK}]$ to construct a color-suppressed counterfactual text. The resulting counterfactual text is denoted as $T^M = \mathcal{I}_c(T)$. The corresponding text embeddings are $C_{txt} = \phi_t(T)$ and $C_{txt}^M = \phi_t(T^M)$. We define the counterfactual color residual as

$$\Delta_c^t = \phi_t(T) - \phi_t(\mathcal{I}_c(T)) = C_{txt} - C_{txt}^M. \tag{7}$$

Here, Δ_c^t estimates the semantic displacement induced by removing color expressions from the textual description, while C_{txt}^M preserves color-suppressed contextual semantics such as clothing category, accessories, hairstyle, and body structure.

For multispectral visual representations, RGB images contain human-perceived color information, whereas NIR and TIR images mainly preserve color-weak texture, contour, and thermal responses. We therefore define the spectral feature residual as:

$$\Delta_m^v = \phi_v(I^{rgb}) - \phi_v(I^m) = C_{rgb} - C_m, \quad m \in \mathcal{M}^-, \tag{8}$$

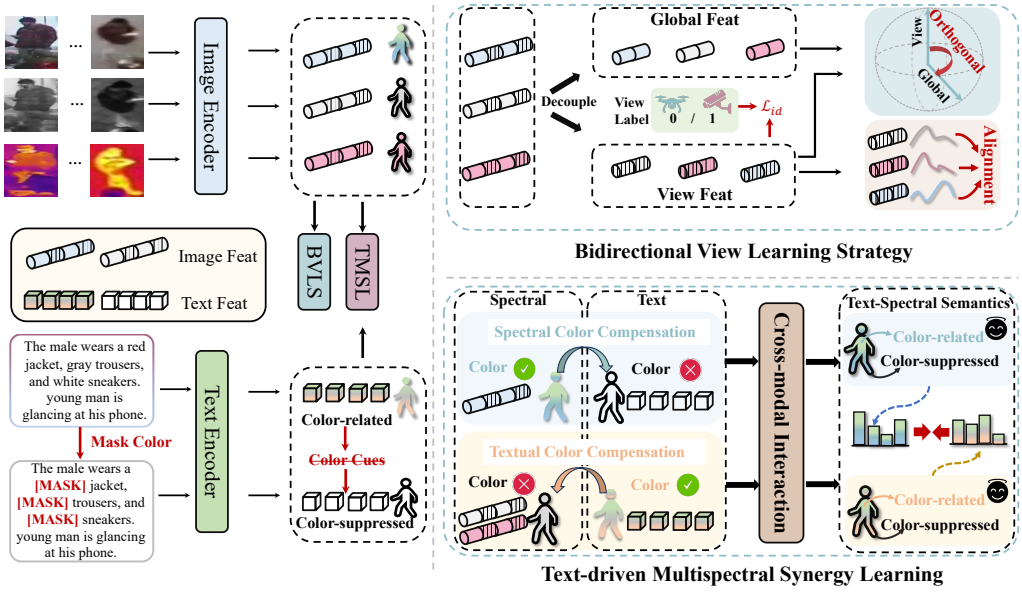

Figure 3: **Overview of our proposed VSRL framework.** It consists of two key components: Bidirectional View Learning Strategy (BVLS) and Text-driven Multispectral Synergy Learning (TMSL). BVLS reduces the reliance of global features on view-related cues within each modality. It also enhances cross-modal consistency of view representations to learn more robust visual features. TMSL conducts bidirectional color compensation: RGB color cues enrich color-suppressed textual features, while textual color residuals compensate color-weak NIR/TIR spectral features. The compensated features are fed into the Cross-modal Interaction module to produce robust text-spectral semantics containing both color-related and color-suppressed information for cross-modal alignment.

where $\mathcal{M}^- = \{nir, tir\}$ represents the color-weak spectra. Although Δ_m^v is not a pure color vector, it contains the dominant feature displacement from RGB to the color-weak spectrum m, including the RGB-specific color discrepancy.

The core assumption of TMSL is that the textual color residual and the RGB-to-NIR/TIR spectral displacement share a dominant color-related residual subspace. We conduct validation experiments and the detailed experimental results presented in Section 5.4 support the above assumption. We denote this text-spectral residual correspondence as:

$$\Delta_c^t \overset{\mathcal{S}_c}{\sim} \Delta_m^v \iff \mathrm{Proj}_{\mathcal{S}_c}(\Delta_c^t) \approx \mathrm{Proj}_{\mathcal{S}_c}(\Delta_m^v), \quad m \in \mathcal{M}^-, \tag{9}$$

where $\mathcal{S}_c$ denotes the latent color-related residual subspace, and $\mathrm{Proj}_{\mathcal{S}_c}(\cdot)$ represents the conceptual projection onto this subspace. However, directly matching Δ_c^t and Δ_m^v is vulnerable to sensor-specific noise, illumination variation, and non-color spectral discrepancies. Therefore, instead of enforcing hard residual regression, we transform the residual correspondence into a residual-compensated composition problem. We define a unified residual compensation operator:

$$\hat{F}_m = \mathcal{R}_m(C_m, T) = C_m + C_{txt}^M + \gamma_m \Delta_c^t, \quad m \in \{rgb, nir, tir\}, \tag{10}$$

where

$$\gamma_m = \begin{cases} 0, & m = rgb, \\ 1, & m \in \mathcal{M}^-. \end{cases} \tag{11}$$

Accordingly, this formulation gives three residual-compensated text-spectral representations. The RGB branch $\hat{F}_{rgb}$ already contains color cues and is regularized by the color-suppressed textual context C_{txt}^M. In contrast, the color-weak NIR and TIR branches $\hat{F}_{nir}/\hat{F}_{tir}$ are compensated by the textual color residual Δ_c^t, enabling them to recover missing color-sensitive semantics from text.

For any $m \in \mathcal{M}^-$, we have

$$\hat{F}_m - \hat{F}_{rgb} = \left(C_m + C_{txt}^M + \Delta_c^t\right) - \left(C_{rgb} + C_{txt}^M\right) = \Delta_c^t - \Delta_m^v. \tag{12}$$

We term this relation residual closure because the discrepancy between $\hat{F}_{nir}/\hat{F}_{tir}$ and $\hat{F}_{rgb}$ is expressed as the difference between the textual color residual and the RGB-to-m visual residual. This relation provides a composition-based way to model text-spectral residual correspondence, rather than requiring direct regression between heterogeneous residuals.

To incorporate the residual-closure relation during training, we use the residual-compensated representation $\hat{F}_m$ as a text-spectral anchor to guide patch-level visual learning. Specifically, $\hat{F}_m$ concatenated with the corresponding patch-level visual tokens are fed into a Cross-modal Interaction Network:

$$\widetilde{F}_m = \mathcal{G}_{CIN}\left(Concat(\hat{F}_m; F_m^1; F_m^2; \cdots; F_m^l)\right), \quad m \in \{rgb, nir, tir\}, \tag{13}$$

where $\mathcal{G}_{CIN}$ denotes the Cross-modal Interaction Network, consisting of a multi-head cross-attention (MCA) layer and four-layer transformer blocks. Through this design, the residual-compensated anchor guides each spectral branch to selectively reorganize its patch-level pedestrian cues under textual semantic supervision.

Then we enforce consistency of the deeply fused features via the following loss function:

$$\mathcal{L}_{TMSL}^{m2n} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\langle\widetilde{F}_m^i, \widetilde{F}_n^i\rangle/\tau)}{\sum_{j=1}^{B}\exp(\langle\widetilde{F}_m^i, \widetilde{F}_n^j\rangle/\tau)}, \tag{14}$$

where $m, n \in \{rgb, nir, tir\}$, τ is a temperature parameter to control the degree of discrimination, and B represents the size of a mini-batch. $\langle\cdot,\cdot\rangle$ denotes inner product.

Symmetrically, we formulate the TMSL loss as:

$$\mathcal{L}_{TMSL} = \sum_{m,n\in\{rgb,nir,tir\}}^{m\neq n}(\mathcal{L}_{TMSL}^{m2n} + \mathcal{L}_{TMSL}^{n2m}). \tag{15}$$

This objective encourages different spectral branches of the same identity to reach agreement after residual-compensated composition, while separating mismatched identities within the batch. Finally, the overall objective for training is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_b + \alpha\mathcal{L}_{BVLS} + \beta\mathcal{L}_{TMSL}, \tag{16}$$

where α and β are the hyperparameters to balance the contribution of each loss function in $\mathcal{L}_{total}$.

5 Experiments

5.1 Experimental Settings

Evaluation Protocols. We follow standard practice and adopt Rank-k metrics (k=1, 5, 10) as the primary evaluation criteria. Specifically, Rank-k measures the likelihood that at least one correctly matched image appears within the top-k retrieved results given a textual query. To provide a more comprehensive assessment of retrieval performance, we further report mean Average Precision (mAP), which captures the overall retrieval accuracy. Higher values indicate better performance.

Implementation Details. Our implementation is based on PyTorch with an NVIDIA 3090 GPU. We utilize the pre-trained CLIP-ViT-B/16 and CLIP text Transformer as the image and text encoder respectively. Images are resized to 384×128 and augmented with random flipping, cropping, and erasing. We train for 50 epochs with Adam, batch size 32, text length 77, and learning rate 5×10^{-6}. The hyperparameters are set to $\tau = 0.02$, $\alpha = 1.0$, and $\beta = 0.5$. At inference time, pedestrian retrieval is achieved by measuring the similarity between the textual embedding and the fused representation formed by summing the features from three spectral modalities.

5.2 Comparison with State-of-the-art Methods

Table 1 compares VSRL with state-of-the-art text-based ReID methods [15, 18, 3, 29, 39]. Since existing methods target single RGB images, we adapt them to TMAG by applying their visual encoders to each spectral modality and summing the tri-spectral features. All models are trained on the full aerial-ground training set and evaluated under mixed, ground-only, and aerial-only settings.

Table 1: Comparison with the state-of-the-art methods on TMAG. Each model is trained on the complete training set that jointly includes both ground and aerial views. Then we evaluate all models on the full, ground-view-only and aerial-view-only test set, respectively.

Methods	*TMAG*				*TMAG (ground)*				*TMAG (aerial)*			
	Rank-1	*Rank-5*	*Rank-10*	*mAP*	*Rank-1*	*Rank-5*	*Rank-10*	*mAP*	*Rank-1*	*Rank-5*	*Rank-10*	*mAP*
IRRA [15]	28.77	48.63	59.59	20.92	30.38	50.28	62.79	21.98	22.92	38.36	52.63	19.52
AUL [18]	28.13	46.80	57.53	20.46	29.54	50.91	61.87	21.23	23.19	41.32	51.83	19.71
RDE [29]	29.68	49.32	60.73	20.91	32.13	52.62	63.44	22.30	23.06	45.21	55.48	19.66
TBPS-CLIP [3]	29.37	50.17	61.87	21.32	31.46	52.14	63.35	22.09	22.60	44.29	56.16	19.78
AEA-FIRM [39]	29.07	48.73	60.05	20.73	30.91	51.47	61.85	22.05	24.08	45.59	56.39	20.12
VSRL (Ours)	**32.42**	**52.51**	**63.70**	**22.03**	**34.47**	**53.43**	**65.07**	**23.06**	**27.85**	**50.60**	**60.47**	**20.76**

Table 2: Ablation results for each component on TMAG are reported on the mixed aerial-ground and the single-view test sets respectively.

Components		*TMAG*				*TMAG (ground)*				*TMAG (aerial)*			
BVLS	*TMSL*	*Rank-1*	*Rank-5*	*Rank-10*	*mAP*	*Rank-1*	*Rank-5*	*Rank-10*	*mAP*	*Rank-1*	*Rank-5*	*Rank-10*	*mAP*
		27.63	47.95	57.76	20.58	28.08	49.77	62.33	21.10	21.28	37.94	47.52	19.25
✔		29.91	49.54	61.19	21.12	31.37	52.17	63.13	21.91	25.34	45.66	58.45	20.61
	✔	30.17	50.26	62.31	21.25	32.06	52.58	64.79	22.30	26.83	47.92	59.03	20.65
✔	✔	**32.42**	**52.51**	**63.70**	**22.03**	**34.47**	**53.43**	**65.07**	**23.06**	**27.85**	**50.60**	**60.47**	**20.76**

VSRL consistently achieves the best performance, surpassing the second-best method by 2.74% Rank-1 on the mixed set, at least 2.34% on the ground-only set, and 3.77% on the challenging aerial-only set. These results validate the effectiveness of VSRL in handling viewpoint variations and exploiting complementary multispectral cues.

5.3 Ablation Studies

Table 2 demonstrates the complementary benefits of BVLS and TMSL across aerial-ground, ground-only, and aerial-only settings. On the mixed test set, BVLS improves Rank-1 from 27.63 to 29.91, while TMSL raises it to 30.17 via counterfactual color residual compensation. Their combination achieves the best Rank-1/mAP scores of 32.42/22.03, 34.47/23.06, and 27.85/20.76, confirming the effectiveness of jointly modeling view invariance and tri-spectral text-image alignment.

Visualization. To visualize the effects of BVLS and TMSL, we show t-SNE [36] feature distributions in Figure 4, where triangles and squares denote text and summed tri-spectral image features, respectively, and colors indicate identities. In the baseline, text and image features of the same identity are scattered and poorly aligned. With BVLS and TMSL, intra-identity text-image distances are reduced and inter-identity separation becomes clearer, demonstrating their effectiveness in improving cross-modal feature alignment.

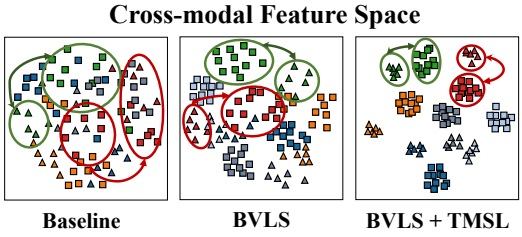

Figure 4: t-SNE visualization. Triangles denote text features, squares denote tri-spectral image features, and colors indicate identities.

5.4 Color-Residual Alignment Validation

The core assumption in Section 4.3 is that the RGB-to-NIR/TIR visual residual and the counterfactual textual color residual share a dominant color-related residual subspace. To empirically examine the assumption, we conduct three complementary analyses on TMAG. To avoid circular validation, all features in this subsection are extracted from the *Baseline+BVLS* checkpoint before applying *TMSL*.

Identity-aware Alignment Analysis. We first examine whether the textual color residual selectively aligns with the visual spectral discrepancy by computing the cosine similarities between $\Delta^v_{nir}/\Delta^v_{tir}$ and C_{txt}/Δ^t_c. C^+_{text} and Δ^{t+}_c denote features from the same identity, while C^-_{text} and Δ^{t-}_c denote features from different identities. As shown in Figure 5(a), Δ^t_c yields stronger matched alignment

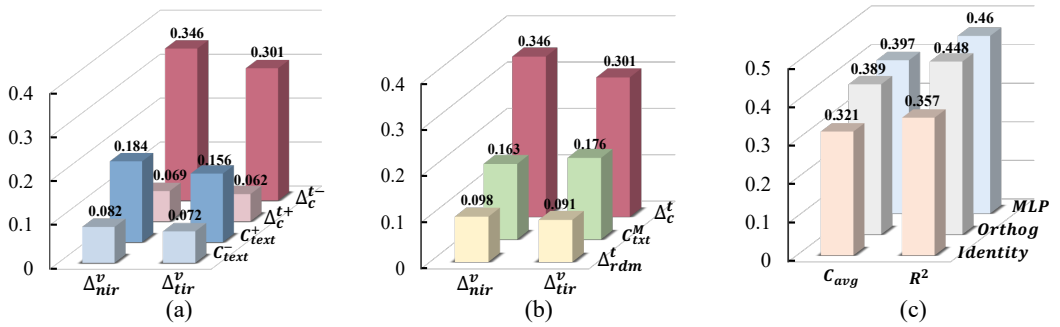

Figure 5: Validation of the color-residual alignment hypothesis in Eq. (9). (a) Δ_c^t aligns better with $\Delta_{nir}^v/\Delta_{tir}^v$ than C_{txt}. (b) Removing or perturbing color words weakens the alignment, confirming its color-related nature. (c) An orthogonal linear map substantially improves recoverability over the identity mapping, while a two-layer MLP brings only marginal further gains.

than C_{txt} for both spectra: the similarity increases from 0.184 to 0.346 for Δ_{nir}^v and from 0.156 to 0.301 for Δ_{tir}^v, while mismatched similarities remain low or decrease. These results show that the visual spectral discrepancy aligns more strongly with the textual color-related residual than the raw text embedding, indicating that isolating color-related semantics yields a more discriminative color-related cue for text-spectral alignment.

Counterfactual Color Intervention. We further examine whether the text-spectral alignment is induced by semantically correct color cues. We compare $\Delta_{nir}^v/\Delta_{tir}^v$ with the original color residual Δ_c^t, the color-suppressed context C_{txt}^M, and a random color residual Δ_{rdm}^t obtained by replacing color words with those from other identities. As shown in Figure 5(b), Δ_c^t achieves the strongest alignment, reaching 0.346/0.301, while C_{txt}^M obtains 0.163/0.176 and Δ_{rdm}^t drops to 0.098/0.091. This confirms that the residual correspondence is mainly driven by correct color semantics, supporting color-residual compensation in TMSL for color-weak NIR/TIR branches.

Linear Recoverability. We examine whether visual spectral residuals can recover textual color residuals using Identity, Orthogonal, and two-layer MLP mappings. Recoverability is evaluated by Average Cosine Similarity C_{avg} and the coefficient of determination R^2:

$$C_{avg} = \frac{1}{N}\sum_{i=1}^{N}\cos\left(g(x^i), \Delta_c^t(i)\right), R^2 = 1 - \frac{\sum_{i=1}^{N}\|g(x^i) - \Delta_c^t(i)\|_2^2}{\sum_{i=1}^{N}\|\Delta_c^t(i) - \bar{\Delta}_c^t\|_2^2}, \tag{17}$$

where x^i denotes the visual spectral residual, g is the mapping function, and $\bar{\Delta}_c^t$ is the mean textual color residual over the test set. As shown in Figure 5(c), Identity mapping achieves 0.321/0.357 (C_{avg}/R^2), while Orthogonal mapping improves to 0.389/0.448. The MLP yields only marginal gains (0.397/0.460). These results suggest that the dominant correspondence between visual spectral residuals and textual color residuals can be captured by a simple linear transformation, supporting the linearized residual relation and residual-compensated consistency design in Section 4.3.

5.5 Validation of TMAG-ReID Challenges

To investigate whether TMAG-ReID can be effectively addressed by simply combining existing view-focused and spectrum-focused methods, we conduct additional comparisons on TMAG. Specifically, we select TAG-CLIP, which addresses text-based aerial-ground person retrieval, and CSDT, which focuses on text-to-visible-infrared person ReID. We further evaluate their direct combination under the same text-query protocol, using a CLIP-based dual-encoder Transformer as the baseline. As shown in Table 3, TAG-CLIP and CSDT improve the baseline Rank-1/mAP from 27.63%/20.58% to 29.15%/20.78% and 30.08%/20.92%, respectively. However, their direct

Table 3: Comparison with view-focused and spectrum-focused methods and their direct combination on TMAG dataset. All methods are evaluated under the same text-query protocol. Rank-1 and mAP (%) are reported.

Methods	*Rank-1*	*mAP*
Baseline	27.63	20.58
TAG-CLIP [60]	29.15	20.78
CSDT [55]	30.08	20.92
TAG-CLIP + CSDT	29.76	20.81
VSRL (Ours)	**32.42**	**22.03**

combination achieves only 29.76%/20.81%, slightly below CSDT alone. In contrast, VSRL achieves the best performance of 32.42%/22.03%, outperforming CSDT by 2.34% in Rank-1 and 1.11% in mAP. These results suggest that viewpoint and spectral discrepancies jointly create a complex challenge in TMAG-ReID that cannot be adequately addressed by simply combining existing view-focused and spectrum-focused modules.

5.6 Robustness to Missing Spectral Modalities

To evaluate the robustness of VSRL to missing spectral inputs, we train all methods with complete RGB/NIR/TIR observations and evaluate them under three incomplete-modality settings: R+N, R+T, and N+T. During inference, only the two available spectral features are fused and matched with the textual query. As shown in Table 4, VSRL achieves Rank-1/mAP scores of 31.16/21.06, 29.98/20.74, and 26.25/19.97, outperforming the strongest competing methods by 5.87/1.41, 5.02/1.02, and 5.33/3.09 percentage points, respectively. Compared with the complete tri-spectral setting (32.42/22.03), VSRL exhibits relatively small performance degradation when TIR or NIR is missing. A larger decrease is observed without RGB, as textual descriptions are primarily grounded in RGB color attributes. Nevertheless, VSRL consistently achieves the best performance across all three settings, demonstrating its robustness to missing spectral modalities.

Table 4: Comparison with state-of-the-art methods under missing spectral modalities. All methods are trained with complete RGB/NIR/TIR inputs and evaluated using only two available spectral modalities. Results are reported as Rank-1/mAP (%).

Methods	*R + N*	*R + T*	*N + T*
IRRA [15]	23.85 / 19.25	22.34 / 18.72	18.30 / 16.23
AUL [18]	24.08 / 19.11	23.78 / 18.69	19.36 / 16.76
RDE [29]	25.17 / 19.46	24.75 / 19.41	20.13 / 16.82
TBPS-CLIP [3]	25.29 / 19.65	24.96 / 19.72	20.92 / 16.88
AEA-FIRM [39]	24.58 / 19.21	23.87 / 18.85	19.76 / 16.58
VSRL (Ours)	**31.16 / 21.06**	**29.98 / 20.74**	**26.25 / 19.97**

6 Limitations

Despite the improvements achieved by VSRL, the absolute retrieval performance on TMAG remains relatively low. This is mainly attributed to the combined challenges of aerial-ground viewpoint variations, small pedestrian scales, and substantial spectral discrepancies. Moreover, RGB-grounded textual descriptions may contain inaccurate color attributes, which could affect the reliability of textual color residual compensation. The current dataset is also limited in scale, and its template-based descriptions may not fully capture the diversity of real-world language queries. Future work will focus on expanding the dataset, improving robustness to noisy textual descriptions, and enhancing cross-spectral semantic alignment under challenging conditions.

7 Conclusion

In this work, we introduce Text-based Multispectral Aerial-Ground Person Re-Identification (TMAG-ReID), extending text-based ReID to a more realistic multiview-multispectral setting. To support this task, we construct TMAG, the first tri-spectral aerial-ground pedestrian benchmark with reliable textual annotations. We further propose View-Spectral Reconciliation Learning (VSRL) to address viewpoint discrepancy and spectral complementarity. Specifically, BVLS learns view-invariant pedestrian representations, while TMSL reconciles RGB-grounded textual semantics with NIR/TIR spectra through counterfactual color residual compensation. Experiments on TMAG show that VSRL consistently outperforms existing text-based ReID methods, demonstrating the importance of jointly modeling multiview and multispectral cues for robust cross-modal retrieval.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China under Grants (62501428), China Postdoctoral Science Foundation (2026T190427), and Hubei Provincial Natural Science Foundation of China (2025AFB219). The numerical calculations in this paper had been supported by the super-computing system in the Supercomputing Center of Wuhan University.

References

[1] Yang Bai, Jingyao Wang, Min Cao, Chen Chen, Ziqiang Cao, Liqiang Nie, and Min Zhang. Text-based person search without parallel image-text data. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 757–767, 2023.

[2] Yang Bai, Yucheng Ji, Min Cao, Jinqiao Wang, and Mang Ye. Chat-based person retrieval via dialogue-refined cross-modal alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3952–3962, 2025.

[3] Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. An empirical study of clip for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 465–473, 2024.

[4] Cuiqun Chen, Mang Ye, and Ding Jiang. Towards modality-agnostic person re-identification with descriptive query. In *CVPR*, pages 15128–15137, 2023.

[5] Zhenyu Cui, Jiahuan Zhou, Xun Wang, Manyu Zhu, and Yuxin Peng. Learning continual compatible representation for re-indexing free lifelong person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16614–16623, 2024.

[6] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*, 2021.

[7] Takuro Fujii and Shuhei Tarashima. Bilma: Bidirectional local-matching for text-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2786–2790, 2023.

[8] Yunpeng Gong, Liqing Huang, and Lifei Chen. Person re-identification method based on color attack and joint defence. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4312–4321. IEEE, 2022.

[9] Yunpeng Gong, Zhun Zhong, Yansong Qu, Zhiming Luo, Rongrong Ji, and Min Jiang. Cross-modality perturbation synergy attack for person re-identification. *Advances in Neural Information Processing Systems*, 37:23352–23377, 2024.

[10] Yunpeng Gong, Yongjie Hou, Jiangming Shi, Kim Long Diep, and Min Jiang. A theory-inspired framework for few-shot cross-modal sketch person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 4284–4292, 2026.

[11] Peixian Hong, Tao Wu, Ancong Wu, Xintong Han, and Wei-Shi Zheng. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10513–10522, 2021.

[12] Zhangyi Hu, Bin Yang, and Mang Ye. Empowering visible-infrared person re-identification with large foundation models. *Advances in Neural Information Processing Systems*, 37:117363–117387, 2024.

[13] Yan Huang, Qiang Wu, Zhang Zhang, Caifeng Shan, Yi Zhong, and Liang Wang. Meta clothing status calibration for long-term person re-identification. *IEEE transactions on image processing*, 33:2334–2346, 2024.

[14] Zhipeng Huang, Jiawei Liu, Liang Li, Kecheng Zheng, and Zheng-Jun Zha. Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 1034–1042, 2022.

[15] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2787–2797, 2023.

[16] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[17] Hongchao Li, Jingong Chen, Aihua Zheng, Yong Wu, and Yonglong Luo. Day-night cross-domain vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12626–12635, 2024.

[18] Shenshen Li, Chen He, Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen. Adaptive uncertainty-based learning for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3172–3180, 2024.

[19] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1970–1979, 2017.

[20] Xingfeng Li, Yuangang Pan, Yuan Sun, Quansen Sun, Yinghui Sun, Ivor W. Tsang, and Zhenwen Ren. Incomplete multi-view clustering with paired and balanced dynamic anchor learning. *IEEE Transactions on Multimedia*, pages 7087–7098, 2025.

[21] Xingfeng Li, Hao Pan, Honglin Yuan, Yuan Sun, Xujian Zhao, Jiaqi Lin, and Zhenwen Ren. Anti-degradation lifelong multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8749–8759, 2026.

[22] Yanping Li, Yizhang Liu, Hongyun Zhang, Cairong Zhao, Zhihua Wei, and Duoqian Miao. Occlusion-aware transformer with second-order attention for person re-identification. *IEEE Transactions on Image Processing*, 33:3200–3211, 2024.

[23] Min Liu, Fei Wang, Xueping Wang, Yaonan Wang, and Amit K Roy-Chowdhury. A two-stage noise-tolerant paradigm for label corrupted person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4944–4956, 2024.

[24] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

[25] Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Ag-reid. v2: Bridging aerial and ground views for person re-identification. *IEEE Transactions on Information Forensics and Security*, 19:2896–2908, 2024.

[26] Vuong D Nguyen, Samiha Mirza, Abdollah Zakeri, Ayush Gupta, Khadija Khaldi, Rahma Aloui, Pranav Mantini, Shishir K Shah, and Fatima Merchant. Tackling domain shifts in person re-identification: A survey and analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4149–4159, 2024.

[27] Hao Pan, Jian Dai, Yuan Sun, Zhenwen Ren, and Xingfeng Li. Difference-aware decision learning for multimodal image fusion. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 95745–95762. PMLR, 2026.

[28] Honghu Pan, Wenjie Pei, Xin Li, and Zhenyu He. Unified conditional image generation for visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 2024.

[29] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27197–27206, 2024.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[31] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[32] Jiangming Shi, Xiangbo Yin, Yeyun Chen, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Multi-memory matching for unsupervised visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 456–474. Springer, 2024.

[33] Yuan Sun, Zhenwen Ren, Peng Hu, Dezhong Peng, and Xu Wang. Hierarchical consensus hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 26:824–836, 2023.

[34] Yuan Sun, Yang Qin, Yongxiang Li, Dezhong Peng, Xi Peng, and Peng Hu. Robust multi-view clustering with noisy correspondence. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):9150–9162, 2024.

[35] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.

[36] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[38] Changshuo Wang, Shuting He, Meiqing Wu, Siew-Kei Lam, Prayag Tiwari, and Xingyu Gao. Looking clearer with text: A hierarchical context blending network for occluded person re-identification. *IEEE Transactions on Information Forensics and Security*, 2025.

[39] Yihao Wang, Meng Yang, Rui Cao, and Guangwei Gao. Aea-firm: Adaptive elastic alignment with fine-grained representation mining for text-based aerial pedestrian retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[40] Yuhao Wang, Xuehu Liu, Pingping Zhang, Hu Lu, Zhengzheng Tu, and Huchuan Lu. Top-reid: Multi-spectral object re-identification with token permutation. In *AAAI*, volume 38, pages 5758–5766, 2024.

[41] Yuhao Wang, Xuehu Liu, Tianyu Yan, Yang Liu, Aihua Zheng, Pingping Zhang, and Huchuan Lu. Mambapro: Multi-modal object re-identification with mamba aggregation and synergistic prompt. In *AAAI*, volume 39, pages 8150–8158, 2025.

[42] Yuhao Wang, Yang Liu, Aihua Zheng, and Pingping Zhang. Decoupled feature-based mixture of experts for multi-modal object re-identification. In *AAAI*, volume 39, pages 8141–8149, 2025.

[43] Yuhao Wang, Yongfeng Lv, Pingping Zhang, and Huchuan Lu. Idea: Inverted text with cooperative deformable aggregation for multi-modal object re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29701–29710, 2025.

[44] Zi Wang, Huaibo Huang, Aihua Zheng, and Ran He. Heterogeneous test-time training for multi-modal person re-identification. In *AAAI*, volume 38, pages 5850–5858, 2024.

[45] Yudi Xie, Zhongao Zhou, Bin Yang, Zhenghan Chen, and Mang Ye. Towards cross-modal preservation, consistency and alignment for privacy-preserving visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11273–11282, 2026.

[46] Kunlun Xu, Xu Zou, Yuxin Peng, and Jiahuan Zhou. Distribution-aware knowledge prototyping for non-exemplar lifelong person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16604–16613, 2024.

[47] Bin Yang, Jun Chen, and Mang Ye. Shallow-deep collaborative learning for unsupervised visible-infrared person re-identification. In *CVPR*, pages 16870–16879, 2024.

[48] Honglin Yuan, Yuan Sun, Fei Zhou, Jing Wen, Shihua Yuan, Xiaojian You, and Zhenwen Ren. Prototype matching learning for incomplete multi-view clustering. *IEEE Transactions on Image Processing*, 34:828–841, 2025.

[49] Pingping Zhang, Yuhao Wang, Yang Liu, Zhengzheng Tu, and Huchuan Lu. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *CVPR*, pages 17117–17126, 2024.

[50] Quan Zhang, Jianhuang Lai, Zhanxiang Feng, and Xiaohua Xie. Uncertainty modeling for group re-identification. *International Journal of Computer Vision*, 132(8):3046–3066, 2024.

[51] Quan Zhang, Lei Wang, Vishal M Patel, Xiaohua Xie, and Jianhaung Lai. View-decoupled transformer for person re-identification under aerial-ground camera network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22000–22009, 2024.

[52] Shizhou Zhang, Qingchun Yang, De Cheng, Yinghui Xing, Guoqiang Liang, Peng Wang, and Yanning Zhang. Ground-to-aerial person search: Benchmark dataset and approach. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 789–799, 2023.

[53] Shizhou Zhang, Wenlong Luo, De Cheng, Yinghui Xing, Guoqiang Liang, Peng Wang, and Yanning Zhang. Prompt-based modality alignment for effective multi-modal object re-identification. *IEEE TIP*, 2025.

[54] Yiyuan Zhang, Yuhao Kang, Sanyuan Zhao, and Jianbing Shen. Dual-semantic consistency learning for visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 18:1554–1565, 2022.

[55] Ziang Zhang, Bin Yang, and Mang Ye. Cross-modal semantic decoupling and transfer for text-to-visible-infrared person re-identification. In *Forty-third International Conference on Machine Learning*.

[56] Yuxuan Zhao, Zhongao Zhou, Bin Yang, He Li, Jian Liang, Jun Chen, Bo Du, and Mang Ye. Whu-mars: A multispectral aerial-ground benchmark towards any-scenario person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25461–25471, 2026.

[57] Aihua Zheng, Zi Wang, Zihan Chen, Chenglong Li, and Jin Tang. Robust multi-modality person re-identification. In *AAAI*, volume 35, pages 3529–3537, 2021.

[58] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1367–1376, 2017.

[59] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020.

[60] Xinyu Zhou, Yu Wu, Jiayao Ma, Wenhao Wang, Min Cao, and Mang Ye. Text-based aerial-ground person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29035–29043, 2026.

[61] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM international conference on multimedia*, pages 209–217, 2021.

[62] Chang Zou, Zeqi Chen, Zhichao Cui, Yuehu Liu, and Chi Zhang. Discrepant and multi-instance proxies for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11058–11068, 2023.

[63] Jialong Zuo, Ying Nie, Hanyu Zhou, Huaxin Zhang, Haoyu Wang, Tianyu Guo, Nong Sang, and Changxin Gao. Cross-video identity correlating for person re-identification pre-training. *Advances in Neural Information Processing Systems*, 37:25228–25250, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims are clearly stated in the abstract and the introduction

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We included the discussion of the limitations of our work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: Our work focuses on computer vision applications, not including theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the details of our proposed methods to reproduce the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will release the relevant algorithm code after the paper is accepted.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The training and testing details are available in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to the limited resources, our paper does not provide error bars. Also, previous relative methods do not provide error bars either.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The information of computer resources is provided.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics. The data collection, annotation, and release procedure of TMAG are conducted under responsible research considerations, including privacy protection, research-only use, anonymized submission, and institutional data-release review.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This work may advance multispectral aerial-ground person retrieval, but person ReID raises privacy, surveillance, misuse, and fairness risks. All data are collected with the consent of the recorded individuals, and faces will be blurred before release. TMAG and VSRL are intended only for academic benchmarking.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators and original owners of assets used in the paper are properly credited, and the license and terms of use are explicitly mentioned and properly respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: This paper introduces a new dataset, TMAG. We will provide the data collection protocol, annotation procedure, evaluation split, and experimental details in the paper. The dataset will be released soon.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use an MLLM to generate candidate textual descriptions for the TMAG dataset. The MLLM is used only for candidate annotation generation and is not used to generate experimental results or replace human evaluation. The core method development in this research does not involve LLMs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.